

WHAT IS CLAIMED IS:

1 1. A method of creating an audio-centric audio-visual summary of a video program, said
2 video program having an audio track and an image track, said method comprising:

3 selecting a length of time L_{sum} of said audio-visual summary;

4 examining said audio track and image track;

5 identifying one or more audio segments from said audio track based on one or more
6 predetermined audio, image, speech, and text characteristics which relate to desired content
7 of said audio-visual summary, wherein said identifying is performed in accordance with a
8 machine learning method which relies on previously-generated experience-based learning
9 data to provide, for each of said audio segments in said video program, a probability that a
10 given audio segment is suitable for inclusion in said audio-visual summary;

11 adding said audio segments to said audio-visual summary;

12 performing said identifying and adding in descending order of said probability until
13 the length of time L_{sum} is reached; and

14 selecting only one or more image segments corresponding to the one or more
15 identified audio segments, so as to yield a high degree of synchronization between said one or
16 more audio segments and said one or more image segments.

1 2. A method as claimed in claim 1, wherein said identifying further comprises detecting
2 audio segments comprising non-speech sounds; classifying said non-speech sounds according
3 to contents; and, for each of said non-speech sounds, outputting a starting time code, length,
4 and category.

1 3. A method as claimed in claim 2, wherein, when said audio segments comprise speech,
2 said identifying comprises performing speech recognition on said audio segments to generate

3 speech transcripts, and outputting a starting time code and length for each of said speech
4 transcripts.

1 4. A method as claimed in claim 3, wherein, when there is closed captioning present,
2 said method further comprises aligning the closed captioning and the speech transcripts.

1 5. A method as claimed in claim 4, wherein said identifying further comprises
2 generating speech units either based on said aligning, if said closed captioning is present, or
3 based on said speech transcripts, if said closed captioning is not present, and creating a
4 feature vector for each of said speech units.

1 6. A method as claimed in claim 5, further comprising computing an importance rank for
2 each of said speech units.

1 7. A method as claimed in claim 6, further comprising receiving said speech units and
2 determining identities of one or more speakers.

1 8. A method as claimed in claim 1, wherein said identifying further comprises
2 segmenting said image track into individual image segments.

1 9. A method as claimed in claim 8, further comprising extracting image features and
2 forming an image feature vector for each of said image segments.

1 10. A method as claimed in claim 9, further comprising determining identities of one or
2 more faces for each of said image segments.

1 11. A method as claimed in claim 1, wherein said probability is computed in accordance
2 with a method selected from the group consisting of a Naïve Bayes method, a decision tree
3 method, a neural network method, and a maximum entropy method.

1 12. A method of creating an image-centric audio-visual summary of a video program,
2 said video program having an audio track and an image track, said method comprising:
3 selecting a length of time L_{sum} of said audio-visual summary;
4 examining said image track and audio track of said video program;
5 identifying one or more image segments from said image track based on one or more
6 predetermined image, audio, speech, and text characteristics which relate to desired content
7 of said audio-visual summary, wherein said identifying is performed in accordance with a
8 machine learning method which relies on previously-generated experience-based learning
9 data to provide, for each of said image segments in said video program, a probability that a
10 given image segment is suitable for inclusion in said audio-visual summary;
11 adding said one or more image segments to said audio-visual summary;
12 performing said identifying and adding in descending order of said probability until
13 the length of time L_{sum} is reached; and
14 selecting only one or more audio segments corresponding to the one or more
15 identified image segments, so as to yield a high degree of synchronization between said one
16 or more image segments and said one or more audio segments.

1 13. A method as claimed in claim 12, wherein said identifying comprises segmenting said
2 image track into individual image segments.

1 14. A method as claimed in claim 13, further comprising extracting image features and
2 forming an image feature vector for each of said image segments.

1 15. A method as claimed in claim 14, further comprising determining identities of one or
2 more faces for each of said image segments.

1 16. A method as claimed in claim 12, further comprising selecting a minimum playback
2 time L_{\min} for each of said image segments in said audio-visual summary.

1 17. A method as claimed in claim 16, wherein L_{\min} is sufficiently small relative to
2 L_{sum} such that a relatively large number of audio segments and image segments are provided
3 in said audio-visual summary, to provide a breadth-oriented audio-visual summary.

1 18. A method as claimed in claim 16, wherein L_{\min} is sufficiently large relative to
2 L_{sum} such that a relatively small number of audio segments and image segments are provided
3 in said audio-visual summary, to provide a depth-oriented audio-visual summary.

1 19. A method as claimed in claim 12, wherein said identifying further comprises
2 detecting audio segments comprising non-speech sounds; classifying said non-speech sounds
3 according to contents; and, for each of said non-speech sounds, outputting a starting time
4 code, length, and category.

1 20. A method as claimed in claim 19, wherein, when said audio segments comprise
2 speech, said identifying further comprises performing speech recognition on said audio
3 segments to generate speech transcripts, and outputting a starting time code and length for
4 each of said speech transcripts.

1 21. A method as claimed in claim 20, wherein, when there is closed captioning present,
2 said method further comprises aligning the closed captioning and the speech transcripts.

1 22. A method as claimed in claim 21, wherein said identifying further comprises
2 generating speech units either based on said aligning, if said closed captioning is present, or

3 based on said speech transcripts, if said closed captioning is not present, and creating a
4 feature vector for each of said speech units.

1 23. A method as claimed in claim 22, further comprising computing an importance rank
2 for each of said speech units.

1 24. A method as claimed in claim 23, further comprising receiving said speech units and
2 determining identities of one or more speakers.

1 25. A method as claimed in claim 12, wherein said probability is computed in accordance
2 with a method selected from the group consisting of a Naïve Bayes method, a decision tree
3 method, a neural network method, and a maximum entropy method.

1 26. A method of creating an integrated audio-visual summary of a video program, said
2 video program having an audio track and a video track, said method comprising:

3 selecting a length of time L_{sum} of said audio-visual summary;

4 selecting a minimum playback time L_{min} for each of said image segments to be
5 included in the audio-visual summary;

6 creating an audio summary by selecting one or more desired audio segments until the
7 audio-visual summary length L_{sum} is reached, said selecting being determined in accordance
8 with a machine learning method which relies on previously-generated experience-based
9 learning data to provide, for each of said audio segments in said video program, a probability
10 that a given audio segment is suitable for inclusion in said audio-visual summary;

11 computing, for each of said image segments, a probability that a given image segment
12 is suitable for inclusion in said audio-visual summary in accordance with said machine
13 learning method;

for each of said audio segments that are selected, examining a corresponding image segment to see whether a resulting audio segment/image segment pair meets a predefined alignment requirement;

if the resulting audio segment/image segment pair meets the predefined alignment requirement, aligning the audio segment and the image segment in the pair from their respective beginnings for said minimum playback time L_{min} to define a first alignment point;

repeating said examining and aligning to identify all of said alignment points;

dividing said length of said audio-visual summary into a plurality of partitions, each of said partitions having a time period

either starting from a beginning of said audio-visual summary and ending at the first alignment point; or

starting from an end of the image segment at one alignment point, and ending at a next alignment point; or

starting from an end of the image segment at a last alignment point and ending at the end of said audio-visual summary; and

for each of said partitions, adding further image segments in accordance with the following:

identifying a set of image segments that fall into the time period of that partition;

determining a number of image segments that can be inserted into said partition;

determining a length of the identified image segments to be inserted;

36 selecting said number of the identified image segments in descending order of
37 said probability that a given image segment is suitable for insertion in said audio-
38 visual summary; and

39 from each of the selected image segments, collecting a section from its
40 respective beginning for said time length and adding all the collected sections in
41 ascending time order into said partition.

1 27. A method as claimed in claim 26, wherein said identifying further comprises
2 detecting audio segments comprising non-speech sounds; classifying said non-speech sounds
3 according to contents; and, for each of said non-speech sounds, outputting a starting time
4 code, length, and category.

1 28. A method as claimed in claim 27, wherein, when said audio segments comprise
2 speech, said identifying further comprises performing speech recognition on said audio
3 segments to generate speech transcripts, and outputting a starting time code and length for
4 each of said speech transcripts.

1 29. A method as claimed in claim 28, wherein, when there is closed captioning present,
2 said method further comprises aligning the closed captioning and the speech transcripts.

1 30. A method as claimed in claim 29, further comprising generating speech units either
2 based on said aligning, if said closed captioning is present, or based on said speech
3 transcripts, if said closed captioning is not present, and creating a feature vector for each of
4 said speech units.

1 31. A method as claimed in claim 30, further comprising computing an importance rank
2 for each of said speech units.

1 32. A method as claimed in claim 31, further comprising receiving said speech units and
2 determining identities of one or more speakers.

1 33. A method as claimed in claim 26, wherein L_{\min} is sufficiently small relative to
2 L_{sum} such that a relatively large number of image segments are provided in said audio-visual
3 summary, to provide a breadth-oriented audio-visual summary.

1 34. A method as claimed in claim 26, wherein L_{\min} is sufficiently large relative to
2 L_{sum} such that a relatively small number of image segments are provided in said audio-visual
3 summary, to provide a depth-oriented audio-visual summary.

1 35. A method as claimed in claim 26, wherein said probability that said given audio
2 segment is suitable for inclusion in said audio-visual summary is computed in accordance
3 with a method selected from the group consisting of a Naïve Bayes method, a decision tree
4 method, a neural network method, and a maximum entropy method.

1 36. A method as claimed in claim 26, wherein said probability that said given image
2 segment is suitable for inclusion in said audio-visual summary is computed in accordance
3 with a method selected from the group consisting of a Naïve Bayes method, a decision tree
4 method, a neural network method, and a maximum entropy method.

1 37. A method as claimed in claim 26, wherein said identifying further comprises
2 segmenting said image track into individual image segments.

1 38. A method as claimed in claim 37, further comprising extracting image features and
2 forming an image feature vector for each of said image segments.

1 39. A method as claimed in claim 38, further comprising determining identities of one or
2 more faces for each of said image segments.

1 40. A method of creating an audio-centric audio-visual summary of a video program, said
2 video program having an audio track and an image track, said method comprising:

3 selecting a length of time L_{sum} of said audio-visual summary;

4 examining said audio track and image track;

5 identifying one or more audio segments from said audio track based on one or more
6 predetermined audio, image, speech, and text characteristics which relate to desired content
7 of said audio-visual summary, wherein said identifying is performed in accordance with a
8 predetermined set of heuristic rules to provide, for each of said audio segments in said video
9 program, a ranking so as to determine whether a given audio segment is suitable for inclusion
10 in said audio-visual summary;

11 adding said audio segments to said audio-visual summary;

12 performing said identifying and adding in descending order of said ranking of audio
13 segments until the length of time L_{sum} is reached; and

14 selecting only one or more image segments corresponding to the one or more
15 identified audio segments, so as to yield a high degree of synchronization between said one or
16 more audio segments and said one or more image segments.

1 41. A method as claimed in claim 40, wherein said identifying further comprises
2 detecting audio segments comprising non-speech sounds; classifying said non-speech sounds
3 according to contents; and, for each of said non-speech sounds, outputting a starting time
4 code, length, and category.

1 42. A method as claimed in claim 41, wherein, when said audio segments comprise
2 speech, said identifying comprises performing speech recognition on said audio segments to
3 generate speech transcripts, and outputting a starting time code and length for each of said
4 speech transcripts.

1 43. A method as claimed in claim 42, wherein, when there is closed captioning present,
2 said method further comprises aligning the closed captioning and the speech transcripts.

1 44. A method as claimed in claim 43, further comprising generating speech units either
2 based on said aligning, if said closed captioning is present, or based on said speech
3 transcripts, if said closed captioning is not present, and creating a feature vector for each of
4 said speech units.

1 45. A method as claimed in claim 44, further comprising receiving said speech units and
2 determining identities of one or more speakers.

1 46. A method as claimed in claim 40, wherein said identifying comprises segmenting said
2 image track into individual image segments.

1 47. A method as claimed in claim 46, further comprising extracting image features and
2 forming an image feature vector for each of said image segments.

3 48. A method as claimed in claim 47, further comprising determining identities of one or
4 more faces for each of said image segments.

1 49. A method as claimed in claim 40, further comprising computing said ranking for each
2 of said speech units.

1 50. A method of creating an image-centric audio-visual summary of a video program,
2 said video program having an audio track and an image track, said method comprising:

3 selecting a length of time L_{sum} of said summary;

4 examining said image track and audio track;

5 identifying one or more image segments from said image track based on one or more
6 predetermined image, audio, speech, and text characteristics which relate to desired content
7 of said audio-visual summary, wherein said identifying is performed in accordance with a
8 predetermined set of heuristic rules to provide, for each of said image segments in said video
9 program, a ranking so as to determine whether a given image segment is suitable for
10 inclusion in said audio-visual summary;

11 adding said one or more image segments to said audio-visual summary;

12 performing said identifying and adding in descending order of said ranking until the
13 length of time L_{sum} is reached; and

14 selecting only one or more audio segments corresponding to the one or more
15 identified image segments, so as to yield a high degree of synchronization between said one
16 or more image segments and said one or more audio segments.

1 51. A method as claimed in claim 50, wherein said identifying comprises clustering
2 image segments of said video program based on predetermined visual similarity and dynamic
3 characteristics.

1 52. A method as claimed in claim 51, wherein said identifying comprises segmenting said
2 image track into individual image segments.

1 53. A method as claimed in claim 52, further comprising extracting image features and
2 forming an image feature vector for each of said frame clusters.

1 54. A method as claimed in claim 53, further comprising determining identities of one or
2 more faces for each of said frame clusters.

1 55. A method as claimed in claim 50, wherein said identifying further comprises
2 detecting audio segments comprising non-speech sounds, classifying said non-speech sounds
3 according to contents; and, for each of said non-speech sounds, outputting a starting time
4 code, length, and category.

1 56. A method as claimed in claim 55, wherein, when said audio segments comprise
2 speech, said identifying comprises performing speech recognition on said audio segments to
3 generate speech transcripts, and outputting a starting time code and length for each of said
4 speech transcripts.

1 57. A method as claimed in claim 56, wherein, when there is closed captioning present,
2 said method further comprises aligning the closed captioning and the speech transcripts.

1 58. A method as claimed in claim 57, further comprising generating speech units either
2 based on said aligning, if said closed captioning is present, or based on said speech
3 transcripts, if said closed captioning is not present, and creating a feature vector for each of
4 said speech units.

1 59. A method as claimed in claim 58, further comprising computing an importance rank
2 for each of said speech units.

1 60. A method as claimed in claim 59, further comprising receiving said speech units and
2 determining identities of one or more speakers.

61. A method as claimed in claim 50, further comprising selecting a minimum playback time L_{\min} for each of said image segments in said audio-visual summary.

62. A method as claimed in claim 61, wherein L_{\min} is sufficiently small relative to L_{sum} such that a relatively large number of audio segments and image segments are provided in said audio-visual summary, to provide a breadth-oriented audio-visual summary.

63. A method as claimed in claim 61, wherein L_{\min} is sufficiently large relative to L_{sum} such that a relatively small number of audio segments and image segments are provided in said audio-visual summary, to provide a depth-oriented audio-visual summary.

64. A method of creating an integrated audio-visual summary of a video program, said video program having an audio track and a video track, said method comprising:

selecting a length L_{sum} of said audio-visual summary;

selecting a minimum playback time L_{\min} for each of a plurality of image segments to be included in the audio-visual summary;

creating an audio summary by selecting one or more desired audio segments, said selecting being determined in accordance with a predetermined set of heuristic rules to provide, for each of said audio segments in said video program, a ranking to determine whether a given audio segment is suitable for inclusion in said video summary;

performing said selecting in descending order of said ranking of audio segments until said audio-visual summary length is reached;

grouping said image segments of said video program into a plurality of frame clusters based on a visual similarity and a dynamic level of said image segments, wherein each frame

cluster comprises at least one of said image segments, with all the image segments within a given frame cluster being visually similar to one another;

for each of said audio segments that are selected, examining a corresponding image segment to see whether a resulting audio segment/image segment pair meets a predefined alignment requirement;

if the resulting audio segment/image segment pair meets the predefined alignment requirement, aligning the audio segment and the image segment in the pair from their respective beginnings for said minimum playback time L_{\min} to define a first alignment point;

repeating said examining and aligning to identify all of said alignment points;

dividing said length of said audio-visual summary into a plurality of partitions, each of said partitions having a time period

either starting from a beginning of said audio-visual summary and ending at the first alignment point; or

starting from an end of the image segment at one alignment point, and ending at a next alignment point; or

starting from an end of the image segment at a last alignment point and ending at the end of said audio-visual summary; and

dividing each of said partitions into a plurality of time slots, each of said time slots having a length equal to said minimum playback time L_{\min} ;

assigning said frame clusters to fill said time slots of each of said partitions based on the following:

assigning each frame cluster to only one time slot; and

maintaining a time order of all image segments in the audio-visual summary;

37 wherein said assigning said frame clusters to fill said time slots is performed in
38 accordance with a best matching between said frame clusters and said time slots.

1 65. A method as claimed in claim 64, wherein said best matching is computed by a
2 method of maximum-bipartite-matching.

1 66. A method as claimed in claim 65, wherein, if there are more time slots than frame
2 clusters, identifying those frame clusters which contain more than one image segment, and
3 assigning image segments from said identified frame clusters to time slots until all of said
4 time slots are filled, while maintaining said time order of said image segments in said audio-
5 visual summary.

1 67. A method as claimed in claim 66, further comprising reviewing said audio-visual
2 summary to ensure that said time order is maintained, and, if said time order is not
3 maintained, reordering said image segments that were added in each partition so that said
4 time order is maintained.

1 68. A method as claimed in claim 64, wherein said identifying further comprises
2 detecting audio segments comprising non-speech sounds, classifying said non-speech sounds
3 according to contents; and, for each of said non-speech sounds, outputting a starting time
4 code, length, and category.

1 69. A method as claimed in claim 68, wherein, when said audio segments comprise
2 speech, said identifying comprises performing speech recognition on said audio segments to
3 generate speech transcripts, and outputting a starting time code and length for each of said
4 speech transcripts.

- 1 78. A method as claimed in claim 77, further comprising determining identities of one or
- 2 more faces for each of said image segments.